AD_____

Award Number: DAMD17-02-1-0416

TITLE: Novel Targets for Diagnosis and Treatment of Breast Cancer Identified by Genomic Analysis

PRINCIPAL INVESTIGATOR: Carol A. Westbrook, M.D., Ph.D.

CONTRACTING ORGANIZATION: University of Illinois
Chicago, IL 60612-7205

REPORT DATE: May 2004

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 2004	Annual (1 May 2003 - 30 Apr 2004)

4. TITLE AND SUBTITLE

Novel Targets for Diagnosis and Treatment of Breast Cancer Identified by Genomic Analysis

5. FUNDING NUMBERS

DAMD17-02-1-0416

6. AUTHOR(S)

Carol A. Westbrook, M.D., Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Illinois
Chicago, IL 60612-7205

E-Mail: cwcw@uic.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

Chromosomal rearrangements which result in localized increases of genetic material are frequent in breast cancer, and occur consistently in certain genomic regions. The resulting increase in expression of genes contained within these amplifications contributes to the malignant phenotype. Such amplified genes, such as Her2-Neu, provide targets for diagnosis and for the development of inhibitory drugs. The purpose of this study is to use novel genomic technologies to find new genes in breast cancer that are both highly amplified and are suitable targets by virtue of the fact that they are membrane-associated (receptors, membrane antigens, secretory proteins). The aims of the study are (1) To specify intervals of genomic amplification (amplicons) in primary breast cancer cell lines using genomic microarrays; (2) To prepare a database of membrane-associated genes, selected by differential hybridization of RNA prepared from fractionated microsomes; (3) To use the database to select membrane-associated genes that are located within amplicons, and measure their expression in the primary cell line using cDNA arrays, in order to select those that are upregulated. These genes will provide new insights and reagents for diagnosis and treatment of breast cancer.

14. SUBJECT TERMS

Genomic amplification, Amplicons, Genomic Arrays; Expression Arrays; Gene Discovery; Diagnostic Markers; Array CGH

15. NUMBER OF PAGES

20

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

BODY

I. Progress on Task 1.

1. Task 1: To specify intervals of genomic amplification in breast cancer cell lines using genomic microarrays

a). to analyze 16 primary breast cancer cell lines with a 1-Mb resolution genomic microarray

(Month 1-20)

b). to analyze the hybridization data for amplicon boundaries, minimal interval and amplicon

peaks, in order to specify the genomic sequence positions of at least 25 recurrent breast cancer amplicons (Month 20-24)

Aim 1a): In progress (8 cell lines completed on 1-Mb array)

Aim 1b: Not completed.

We have completed a preliminary study of eight breast cancer cell lines, established by our collaborator Dr. R. Mehta. All cell lines had been developed from patient specimens of infiltrating ductal carcinoma collected at the University of Illinois. The tissue was taken from either a (1) resected breast primary (for the primary cell line group) or(2) pleural fluid in patients with pleural metastases (the metastatic cell line group).

2.Methods and Results, Task 1.

DNA copy number analysis was performed using competitive DNA hybridization on glass slides.

Our initial plan was to use Spectral Genomics 1-Mb microarrays; however, as we reported in the previous progress report, the 1-Mb arrays had not been produced by Spectral Genomics as of last year.

This year we tested their lower-resolution arrays (3 – 5 Mb) with unsatisfactory results—we were unable to obtain an adequate comparative hybridization pattern using a control cell line (data not shown).

Thus, we settled on the Vysis Genosensor chip, which contains probes representing genetic segments that are frequent sites of genomic amplification and genomic loss in cancer, with an approximate 3 Mb resolution. Only 7 cell lines were viable after thwaing; a total of $1x\ 10^7$ cells was harvested and DNA prepared from each cell line. DNA amplification analysis was performed using the GenoSensor Array 300 system (Vysis, Downers Grove, IL) as follows. Equal portions of sample DNA labeled with Cy-3 fluor and reference DNA (female DNA obtained from Stratagene) labeled with Cy-5 fluor were hybridized to the GenoSensor Array 300 for 48 hours according to the manufacturer's specifications.

The chips were scanned and analyzed at Vysis Inc. Dr. Suneel Mundle at Vysis kindly provided this
service since a scanner was not available to us. The probes were scored for sites in which there was a
statistical difference compared to normal DNA, that is, increased or decreased (2-fold).
Not unexpectedly, when compared as a group, there were no statistically significant amplifications that distinguished the metastatic group from the primary group. Of note 10p11-15 is a region which appears
to be highly amplified in one of the pleural metastatic lines (BCA-1) and moderately amplified in another metastatic line (BCA-4). 10p11-15 shows no amplification in the cell lines derived from primary sites.
This interval had not previously been noted to be a site of amplification in breast cancer. Data on the
other sites is still being processed. This method should allow us to specify the sites of amplification in
all of our breast cancer cell lines, but we recognize that neither the Vysis chip, nor any other that is commercially available at present, provides the resolution that we had hoped, to allow us to more
accurately specify the genomic position of the recurrent breast cancer amplicons.

II Progress on Task 2.

1. **Task 2: To prepare a database containing at most 9,000 to 15,000 genes expressed in the MCF7**

breast cancer cell line that are likely to be membrane-associated.

a). to extract membrane-bound and cytosolic RNA from the breast cancer cell line MCF7 and
evaluate the quality of the separation by real-time PCR (Months 1-4)

b). to generate RNA from each fraction and validate its quality and its derivation from the membrane (Month 5-7)

c). to hybridize each RNA sample to the U95 set of 5 GeneChips and analyze the data to generate
a membrane/cytosolic (m/c) ratio for every expressed gene, selecting those that are most likely
to be membrane-associated (Month 8-12)

d).to prepare the surface-expressed gene database, including UniGene cluster number and sequence position on the genome map (Month 12)

This task has been completed, and the results are being prepared for publication. In the previous year, Tasks 2a – b. were completed. Here we report the completion of aims c and d, which resulted in the preparation of a dataset of membrane-associated and secreted proteins (MAS) in breast cancer and/or on Affymetrix arrays, which includes 3792 genes. We combined a biochemical and bioinformatics approach to identify MAS genes expressed in the MCF-7 breast cancer cell line. Our approach is based on analysis of the differential hybridization levels of RNAs which are physically separated by virtue of their association with polysomes at the endoplasmic reticulum, which is enriched for membrane-associated and secreted proteins; it is specifically applicable to cDNA arrays such as Affymetrix, which utilize single-color

hybridization instead of dual-color comparative hybridizations. Assignment to MAS class membership is based on both the MEM/CYT ratio and an expression threshold, which are calculated empirically from data collected on a reference set of known cytoplasmic and membrane proteins. This method enabled the identification of 810 MAS genes expressed in MCF-7 cells for which this annotation did not exist previously.

2. Methods, Task 2.

2.1 RNA preparation

Total RNA was isolated from pooled sucrose gradient fractions by mixing with TRIzol LS reagent (Invitrogen, Carlsbad, CA) at a 3:1 ratio (3 parts TRIzol to 1 part sucrose), and extracting as recommended by the manufacturer, followed by two additional salt precipitations (0.3M sodium acetate (pH 5.2) with 3 volumes of ethanol).

2.2 Real-time Quantitative Reverse-transcriptase PCR

1st strand cDNA synthesis

For generation of first-strand cDNA, approximately 1 μg of RNA was reverse-transcribed using Superscript II Reverse Transcriptase (RT) Kit in the presence of Oligo (dT) 12-18 in a final 20-μL reaction volume with reverse transcriptase per manufacturer's recommended protocol (Invitrogen) followed by RNase H treatment.

Real-time reverse transcriptase PCR setup

Real time PCR reactions were performed using DNAse-free cDNA templates generated above and SYBR Green PCR Core Reagents (Applied Biosystems, Branchburg, NJ) following manufacturer's protocol with the following modifications: a 25 μl reaction was used which contained 1X SYBR Green PCR mix, 3 mM MgCl2, 1X dNTP blend (0.2 mM of dATP, dCTP, dGTP and 0.4 mM of dUTP), 0.625 U of AmpliTaq Gold, 0.125 U of AmpErase UNG, 50 nM each of forward and reverse primer, and 1 μl of cDNA template. Default PCR amplification cycles were used as specified by the ABI Prism 7700 Sequence Detection System (Applied Biosystems): 50°C for 2 minutes, 95°C for 10 minute, 40 cycles of 15 seconds at 95°C and 60°C for 1 minute. PCR amplification was followed by melting curve analysis using the following 3 hold cycles: 95°C for 15 seconds, 60°C for 20 seconds, and 95°C for 15 seconds, with the ramping time at maximum value 19:59 minutes set at the last hold cycle. PCR amplification analysis was done on Sequence Detector v.1.7a and melting curves were analyzed on Dissociation Curves v.1 according to Applied Biosystems guidelines.

MCF-7 cDNA template was used to generate a relative standard curve for either the endogenous control or the target gene. MCF-7 cDNA was serially diluted at 1:10 dilution factor starting with the highest arbitrary concentration of 50 ng/μl (taken from 1/20th of a reverse transcriptase reaction of 1 μg of starting total RNA). The sample templates were diluted at 1:5. All samples were done in triplicate. The sequence of the primers used in real time RT-PCR is as follows: endogenous control 18S ribosomal RNA (18SRNA): F:5'-GTAACCCGTTGAACCCCATT-3', R:5'-CCATCCAATCGGTAGTAGCG-3' {Schmittgen, 2000 #171} with the expected size of 150 bp; JAM1: F:5'-CCCTCTTGGCTTGATTTTGC-3', R:5'-TGACCTTGACTGATGGCTTC-3' with the expected size of 115 bp. The GAPDH primers were obtained from Applied Biosystems (Foster City, CA). The quantity of target RNA (either JAM1 or GAPDH) was calculated by interpolating from the standard curve generated for

that specific target. Both JAM1 and GAPDH quantities were normalized to 18S rRNA to calculate the relative quantity.

2.2 Microarray hybridization and scanning

10□g of total RNA was labeled, hybridized to Affymetrix U133A microarrays, processed and scanned according to standard Affymetrix protocols. Each experiment was performed in triplicate. The resulting CEL files were processed using the Bioconductor software suite (a set of libraries for R{Ihaka, 1996 #164}). The Robust Multiarray Average (RMA) algorithm{Bolstad, 2003 #166;Irizarry, 2003 #165;Irizarry, 2003 #167} was used for normalization, background correction and expression value calculation.

2.3 Membrane and cytoplasmic gene reference set

A reference set was developed containing genes which are known to have either membrane or cytoplasmic location and are represented on the Affymetrix U133A microarray using an automatic database search based on Swiss-Prot {Boeckmann, 2003 #172}. Each Affymetrix microarray element has a unique Affymetrix Probe ID that can be mapped to at least one SwissProt Accession number[1] . Each Swiss-Prot entry was then searched for "Cellular location" comment tags. Proteins were considered to have MAS localization if the SwissProt "Cellular location" tag contained one of the following identifiers: "Secreted", "Golgi", "Vesicular", "Membrane", "Lysosome", or "Peroxisome". Entries were considered tentative if they contained, "Probable", "Possible", "Potential", and "By similarity" and considered unambiguous if not. Entries that contained "Nuclear", "Nucleus", and "Mitochondrial" were removed as there is some evidence that nuclear and mitochondrial proteins can be synthesized in either pathway. This resulting list was then hand edited to remove entries containing multiple isoforms targeted for different subcellular compartments. Proteins are considered to have Cytoplasmic localization if the Swiss-Prot "Cellular location" tag contained "Cytoplasmic" or "Cytoplasm". Again, entries with "Probable", "Possible", "Potential", and "By similarity" were considered tentative, and entries containing "Nuclear", "Nucleus", and "Mitochondrial" were removed. This list was hand edited to remove any entries with multiple isoforms as well as entries that contained any references to membrane association or organelles.

2.7 Statistical calculations

At a given MEM/CYT expression ratio r, the probability of belonging to the membrane class ($p(m|r)$) is calculated by using Bayes' rule as shown in Equation 1.

$$p(m \mid r) = \frac{p(r \mid m)P_m}{p(r \mid m)P_m + p(r \mid c)P_c} \qquad (1)$$

where $p(r|a)$ is the proportion of class a at ratio r. The P_a factor corresponds to the prior probability of belonging to class a. It is reasonable to expect that the prior of belonging to the cytoplasmic class would be much higher than the prior of belonging to the membrane class.

[1]Downloadable at http://www.affymetrix.com/analysis/

Since this calculation has never been done on breast cells (it is possible that this prior would change from tissue to tissue), a prior probability of 0.5 is the best compromise.

Sensitivity is defined as TP/(TP+FP), specificity is defined as TN/(FN+TN), and positive predictive value is defined as (TP/(TP+FN), where TP (true positives) are the number of genes that are labeled correctly, FN (false negatives) are the number of membrane genes that are labeled incorrectly, TN (true negatives) are the number of cytoplasmic genes that are labeled correctly, and FP (false positives) are the number of cytoplasmic genes that are incorrectly labeled.

3. Results of Task 2.

3.1 RNA fractionation and verification

The fractionation of polysomes by sucrose density gradient centrifugation was based on the observation that genes encoding proteins which are membrane-associated or secreted are translated by ribosomes bound to the endoplasmic reticulum (membrane bound polysomes or MBPs), while genes encoding proteins which are cytosolic are translated by ribosomes free in the cytosol (free polysomes or FPs). MBPs, being less dense, rise to the top of the gradient while free polysomes remain near the bottom of the gradient.

Here, the method was used to separate intact polysomes prepared from the MCF-7 breast cancer cell line. After centrifugation, the RNA content of successive fractions of the sucrose gradient was estimated by UV spectrophotometry. As seen in Figure 1, the separation resulted in two distinct peaks of absorbance at 260 nm, with the middle fractions having near-baseline absorption. Fractions from each peak were pooled, and the peak nearest the bottom (2 though 15) of the gradient are designated cytoplasmic (CYT), and fractions in the peak nearest the top of the gradient (20 though 26) are designated membrane and secreted (MEM). Sucrose fractions with near-baseline absorption (16 though 19) in between the CYT and MEM pools were saved as negative controls. The peak at the surface of the gradient (the top 1.5 ml fraction) was discarded.

To confirm that the MEM and CYT pools were enriched for MAS-associated and cytoplasmic-associated mRNA, respectively, real-time quantitative reverse-transcriptase PCR (RTqPCR) was performed using two primer pairs expected to amplify coding sequences specific for each population. Junctional adhesion molecule (JAM1) is primarily cell surface-associated, while glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a protein found free in the cytoplasm. Due to their different biological sequestering, we expected JAM11to be more highly-represented in the MBP RNA, while the opposite will be true for GAPDH. To confirm the physical separatioin of these two RNAs, the MEM/CYT expression ratio was calculated (see Table 1). As seen in Table 1, the MEM/CYT expression ratio is about a thousand-fold greater for JAM1 than for GAPDH, demonstrating a marked enrichment in the MEM pool for JAM1.

The RNA pools were then labelled and hybridized to Affymetrix U133A microarrays. MEM expression and CYT expression are the normalized, un-log-transformed value measured on the Affymetrix array hybridized to MEM and CYT RNA, respectively, and were calculated for each microarray element by averaging the expression value across the appropriate triplicate. The MEM/CYT ratio for GAPDH (AFFX-HUMGAPDH/M33197_M_at) and JAM1 (221664_s_at) was then calculated, as shown in Table 1. As expected, the MEM pool shows an enrichment for JAM1 as compared to GAPDH, while the CYT pool shows an enrichment for GAPDH.

3.2 Reference set construction

Because the distribution of MEM/CYT ratios for either class cannot be known a priori, it was necessary to train a classifier using a reference set of genes with known subcellular localization. Out of 22,283 elements in the Affymetrix U133A array, subcellular location annotation as described in Methods was available for 8,912 elements. Unambiguous MAS annotation was found for 2,932 of these, while unambiguous Cytoplasmic annotation was found for 802 elements. These elements with unambiguous annotation represent the reference set.

3.3 Expression threshold calculation

It is likely that only a subset of the elements on the U133A microarray will be expressed in MCF-7 cells at a level great enough for meaningful measurement. To determine that level, we evaluated our ability to distinguish known MAS genes from known Cytosolic genes in the reference set at varying total expression (E_T) levels, where E_T □ MEM expression + CYT expression. A 10 fold cross- validation was performed, at increasing threshold levels of E_T, including only microarray elements with an E_T value greater than or equal to the threshold. Briefly, for each E_T level, the data was randomly partitioned into 10 groups, using 9 of these groups as a "training" set and the remaining group as a "testing" set. At each E_T level, the MEM/CYT ratio threshold was calculated (as described in Methods) for the training set for that E_T level. The positive predictive value, sensitivity, and specificity were calculated by examining the performance of predicting the testing set for that E_T level, and averages over the 10 groups were recorded. The performance of prediction for E_T thresholds ranging from 22,283 (100% of the microarray elements) to 1,106 (4.9% of microarray elements) was examined. The E_T level that corresponded to the highest sensitivity without a significant drop in positive predictive value or specificity was 633. At this level only 28% of probe sets with the highest E_T. are included, resulting in a final dataset of 6,239 probe sets that pass this threshold filtering. Of those, 531 have unambiguous MAS annotation and 350 are have unambiguous cytoplasmic annotation. Additionally, at this level, our 10 fold cross validation yields a 97.7% positive prediction value with 80% sensitivity and 97% specificity.

3.4 MEM/CYT ratio threshold calculation

All of the 881 probe sets in the reference set with the E_T above the threshold of 633 were used to determine the MEM/CYT ratio that corresponds to the maximum posterior probability of belonging to the MAS class. The distribution of MEM/CYT ratios for genes with known localizations was examined (figure 2) and it is interesting to note that the cytoplasmic genes show a discrete peak, while the MAS genes show a bimodal distribution with a smaller peak that associates with the cytoplasmic genes. An arbitrary cut-off MEM/CYT ratio of 1.08 was selected as giving the maximum posterior probability. Note that above this level, the majority of known cytoplasmic genes are excluded but a sizeable fraction of the MAS genes show a lower MEM/CYT ratio. Thus, genes with a ratio below 1.08 cannot be designated with certainty as either MAS or cytoplasmic.

The distribution of MEM/CYT ratios for the remaining probesets (genes of unknown cellular localization) is plotted in Figure 3. Of these, 810 probesets fall above the expression threshold and above the MEM/CYT ratio of 1.08. These 810 probesets are labeled as "Predicted MAS." The remaining 4034 probesets found above the expression threshold and below the MEM/CYT ratio of 1.08 are labeled as "Indeterminant", because we expect a mixture of cytoplasmic and MAS genes in this range of MEM/CYT ratios. Of the "Predicted MAS" probesets, 343 were found to have a tentative subcellular annotation, but it did not meet the criteria previously established for the reference set. The remaining 467 probesets have no subcellular annotation. A similar percentage of "Indeterminant" probesets were found to have some tentative subcellular annotation (1535 out of 4034).

The SwissProt annotations were searched for terms that might indicate a tentative assignment to a cellular fraction (e.g. "Membrane by similarity" or "nuclear.") Table 2 summarizes the tentative localization annotations for the predicted MAS and Indeterminate groups. Over 70% (244 out of 343) of the predicted MAS probe sets with tentative annotations indicate a MAS subcellular location. Less than 5% (15 out of 343) of the predicted MAS probesets with tentative annotation are thought to be cytoplasmic. The remaining probe sets are thought to be localized to the nucleus, the endoplasmic reticulum, mitochondria and other intracellular organelles Biochemical process annotation was available for these 343 probe sets in Gene Ontology. Over half of these appear to be involved in metabolism, while a third are involved in cell growth. Almost 25% of the predicted MAS class are involved in cell communication. (While these annotations appear to comprise a greater number than the actual number of annotated probe sets, GO is organized in a way such that multiple annotations can correspond to a single probe set.)

In contrast, less than 15% (224 out of 1535) of the Indeterminate probe sets with tentative annotation indicate a MAS localization. Over 27% (425 out of 1535) of these are thought to be cytoplasmic. Interestingly, almost 50% of the indeterminant probesets contain nuclear annotation.

4. Discussion, Task 2.

We describe here a novel set of MAS genes expressed in MCF-7 cells. We are able to annotate 810 genes as membrane-associated or secreted with 97% confidence, of which 467 had no previous subcellular location annotation at all. Two levels of validation strengthen our location predictions. First, we performed 10-fold cross validation which is a robust method for estimating performance on future datasets with similar characteristics. Second, we looked at the tentative annotations of genes which we were able to predict subcellular localization. Our MAS predictions were correct 87% of the time.

This study enabled us to describe a general method of applying density gradient fractionation of mRNA to the Affymetrix platform including a robust statistical analysis. We believe this will be a useful tool for investigators wishing to filter existing or future breast cancer Affymetrix datasets in order to look for MAS genes. Furthermore, we have described an approach that can easily be modified for other tissues or states for comparative studies.

There are a significant number of genes with unambiguous MAS annotation that fall below our MEM/CYT threshold. It is unclear if this is due to a real biological process (some of those MAS genes are not MAS localized in MCF-7 cells, for instance) or a processing artifact. Further experimental analysis is needed to elucidate the mechanism in action. Further study is also needed to determine if the protein localization we discovered for MCF-7 cells holds true when analyzing other breast cancer cells.

The probe sets predicted to be MAS localized appear to have diverse functions as indicated by the Biochemical Process annotation of the Gene Ontology. Biochemical process annotation was available for 343 probe sets, amounting to almost half of the predicted MAS class. Over half of these appear to be involved in metabolism, while a third are involved in cell growth. Almost 25% of the predicted MAS class are involved in cell communication. (While these annotations appear to comprise a greater number than the actual number of annotated probe sets, GO is organized in a way such that many annotations can correspond to a single probe set.)

Progress on Task 3:

Task 3. To select genes within amplicons and measure their expression level, to identify upregulated genes which are membrane-associated or secreted

a. to select membrane-associated genes that are within amplicons, comparing the dataset in Aim 2 with the sequence location of amplicons from Aim 1 and the literature (Month 25)
b. to prepare the cDNA array (Month 25-27)
c. to perform an expression analysis of 16 breast cancer cell lines on the cDNA arrays (Month 28 -33)
d. to analyze the expression data collected here, and survey available databases, to identify highly expressed genes that are contained in amplicons (Month 33-36)

Task 3a. to select membrane-associated genes that are within amplicons, comparing the dataset in Aim 2 with the sequence location of amplicons from Aim 1 and the literature (Month 25).

Results: Task 3a: In progress Our studies in Aim 2 provided a database of 3742 genes which are which are unambiguously specificed as being membrane-associated, including 2932 "known" and 810 "new." We are in the process of comparing these to the literature to ascertain which are likely to be contained within known amplicons. We will also determine which of these are within the amplicons identified in the cell lines in Aim 1a.

Task 3b. to prepare a cDNA array containing MAS genes.

Results, Task 3b: Completed

Although our initial aim was to select membrane-associated genes and use them to prepare a cDNA array, we found instead that the most practical way to proceed was to annotate the genes on Affymetrix arrays, to specify which are MAS genes. Thus we have produced a "virtual"cDNA array, which will enable us to perform our expression analysis of Affymetrix arrays and, more importantly, to analyze published expression databases in breast cancer , expecially those on Affymetrix arrays, to identify MAS genes which have biologic significance. Out of 22,283 elements in the Affymetrix U133A array, existing annotation for 8,912 elements enabled us to specify unambiguos MAS annotation for 2,932 of these; to this we added 810 new MAS genes which are known to be expressed in breast cancer, for a total of 3742 genes. This represents approximately 17% of the elements on the chip. In this way, differential hybridization to Affymetrix chips, either in our lab or in published studies, could be interpreted with an eye toward the identification of MAS gene targets.

Task 3c. to perform an expression study of 16 breast cancer cell lines on the cDNA arrays (Month 28 -33)

Results, Task 3c: In progress. We are in the process of using Affymetrix chips to collect expression data on the 16 cell lines used in Aim 1, to identify highly-expressed genes that are in amplicons. This work is in progress.

Task 3d. to analyze the expression data collected here, and survey available databases, to identify highly expressed genes (Month 33-36).
Results, Task 3d: In Progress
Preliminary results

We used the MAS gene dataset presented in Aim 2 to filter a differential gene expression study in breast cancer which compared tumors with good vs. poor 5-year outcome. (van 't Veer, 2002). We asked whether the MAS localization provided by our study might give additional insight into the interpretation of the results, and allow for the selection of differentially-expressed genes which are membrane localized.

In the van't Veer study, RNA from 98 primary breast tumors was hybridized to cDNA microarrays, and the resultant analysis led to a 231-gene expression profile associated with poor prognosis. It was possible to map 166 of these 231 gene to 269 elements represented on the Affymetrix U133A microarray. Of these 269 elements, 22 were found in our predicted MAS database representing 16 unique genes (see Table 3). Almost half of these (7 out of 16) had no subcellular location annotation in GO or Swissprot, although a number of them had a published characterization. Out of the 10 genes with functional annotation, five are involved in metabolism, two are involved in cell-cycle regulation (including the well known vascular endothelial growth factor (VEGF), also known for its role in angiogenesis), along with one each involved in signal transduction, proteolysis, and calcium binding. It is interesting to note that of the genes without functional annotation, HCCR1 is a putative protooncogene, fucosyltransferase 8 is thought to contribute to malignancy, "G protein-coupled receptor 126" contains a "protein tyrosine phosphatase-like protein" domain, and "hypothetical protein FLJ22341" contains a rhomboid domain, thought to regulate epidermal growth factor receptor expression. Any of these proteins, whose upregulation is associated with poor prognosis in breast cancer, merit further investigation as potential treatment targets.

This study demonstrates the utility of the MAS dataset to analyze published datasets.

KEY RESEARCH ACCOMPLISHMENTS

1. We have prepared the MAS database for breast cancer-associated genes, and also developed a statistical algorithm for the interpretation of similar studies using polysome fractionation.
2. We have used the MAAS dataset to annotate Affymetrix arrays for membrane-associated genes. Out of 22,283 elements in the Affymetrix U133A array, existing annotation for 8,912 elements enabled us to specify unambiguos MAS annotation for 2,932 of these; to this we added 810 new MAS genes which are known to be expressed in breast cancer, for a total of 3742 located on Affymetrix affays.
3. We have demonstrated the use of this annotation to filter published data to select membrane-associated genes which are upregulated in poor-prognosis breast cancer.
4. We have begun preliinary studies using Vysis genomic arrays, demonstrating their suitability for these studies

REPORTABLE OUTCOMES

Published abstract:
Stitziel NO, Mar BG, Liang J, Westbrook CA. Membrane-associated and secreted proteins in the breast cancer transcriptosome. Proceedings of the AACR, Volume 45, Abstract 1655. (2004)

Manuscript in preparation:
Membrane-Associated and Secreted Genes in Breast Cancer , Nathan O. Stitziel, Brenton G. Mar, Jie Liang, and Carol A. Westbrook

CONCLUSIONS

We have made a great deal of progress in the identification of membrane-associated genes which are upregulated in breast cancer. We have accomplished most of our aims within the timeline we specified, with a few modifications based on a smaller number of cell lines, a change in the hybridization platforms (Affymetrix, Vysis), but have already produced some publishable results. We anticipate completion of this work, with the modifications outlined below:

Revised Task 1

a). to complete the analysis of the 16 primary breast cancer cell lines using array CGH on Vysis microarrays (Month 24-36)

b). to analyze the hybridization data and specify the genomic sites of amplification in these 16 cell lines. (Month 24-36).

Revised Task 2: Task completed.

Revised Task 3. To select genes within amplicons and measure their expression level, to identify upregulated genes which are membrane-associated or secreted.

a. to select membrane-associated genes that are within amplicons, comparing the MAS dataset in Aim 2 with the sequence location of amplicons from Aim 1 and the literature (Month 24-36).
b. to perform an expression analysis of breast cancer cell lines on the cDNA arrays (Month 28 –
c. to analyze the expression data collected here, and survey available databases, to identify highly expressed genes (Month 33-36).

APPENDICES

Abstract presented at the AACR meeting

Membrane-Associated and Secreted Proteins in the Breast Cancer Transcriptosome
N. Stitziel, B. Mar, J. Liang, C. Westbrook

Membrane-associated and secreted (MAS) proteins are one of the most important cellular components of a cancer, as they include transmembrane receptors and signaling molecules, adhesion molecules, and secreted autocrine peptides. They are also among the most useful targets for diagnosis and treatment, since their location makes them readily accessible to small molecule inhibitors as well as targeted antibodies. The objective of our study was to establish a database of MAS proteins which are present in a representative breast cancer cell line that could be a starting point for future target identification and validation. Because computational methods for finding MAS genes are limited, we used a biological approach, taking advantage of the fact that MAS proteins are preferentially translated in ribosomes associated with the endoplasmic reticulum, whereas cytosolic proteins are translated freely in the cytosol. Sucrose density centrifugation can be used to fractionate these ribosomes and prepare mRNA enriched for membrane-associated genes (MEM) and cytoplasmic genes (CYT). We used this approach to fractionate RNA from MCF-7 breast cancer cells, and the two fractions were labeled and hybridized in triplicate to Affymetrix U133A oligonucleotide microarrays. Expression ratios (MEM/CYT) were calculated for each of the 22,283 microarray elements by dividing the average signal from the MEM microarrays by the average signal from the CYT microarrays. We used the measured MEM/CYT ratio for 979 genes of known cellular location (772 membrane and 207 cytoplasmic genes) to calculate the probability of MAS class membership using Bayesian statistics, and established that the probabiliy was highest at or above a MEM/CYT ratio of 1.07. Applying this threshold to all of the data collected, we determined that 2,445 Affymetrix elements represent genes expressed in MCF-7 cells that are ≥94% likely to be membrane-associated or secreted, and thus comprise the Breast Cancer MAS (BCMAS) database. Studies are in progress to validate the components of the BCMAS database, and identify useful therapeutic targets including transmembrane kinases, receptors, and circulating peptides. To test the performance of our prediction threshold, 10 fold cross validation was performed. The training data was split randomly into 10 mutually exclusive test subsets. A threshold was calculated for each subset using the remaining training data and performance was calculated using the test subset. An average positive predictive value of 80% was achieved. Further work is underway to expand the training set in order to refine the threshold and increase the predictive value.

Table 1

Target	MEM/CYT ratio, as measured by RTPCR	MEM/CYT ratio, as measured by Affymetrix U133A microarray
GAPDH	0.00064	0.879
JAM11	0.387	3.093

Table 2

Predicted Location	Total probe sets	Probe sets with tentative Subcellular annotation	Tentative MAS annotation	Tentative Cytoplasmic annotation	Tentative Nuclear annotation	Other
MAS	810	343	244 (71.1%)	15 (4.4%)	30 (8.7%)	54 (15.7%)
Indeterminate	4034	1535	224 (14.6%)	425 (27.7%)	751 (48.9%)	135 (8.8%)

TABLE 3.

Affymetrix ID	Accession #	Gene Name	Description	Localization (GO & SwissProt)	MEM/CYT Ratio
212640_at	AF052159		Homo sapiens clone 24416 mRNA sequence	None	1.294415914
212248_at	AK000745		Homo sapiens cDNA FLJ20738 fis, clone HEP08257	None	1.261471012
212250_at	AK000745		Homo sapiens cDNA FLJ20738 fis, clone HEP08257	None	1.231897619
212251_at	AK000745		Homo sapiens cDNA FLJ20738 fis, clone HEP08257	None	1.216957321
201818_at	AF052162	FLJ12443	hypothetical protein FLJ12443	None	1.205331836
218686_s_at	Contig55188_RC	FLJ22341	hypothetical protein FLJ22341	None	1.116226761
219202_at	Contig55188_RC	FLJ22341	hypothetical protein FLJ22341	None	1.133333708
207170_s_at	NM_015416	HCCR1	cervical cancer 1 protooncogene	None	1.080429606
201037_at	D25328	PFKP	phosphofructokinase, platelet	None	1.114811758
219197_s_at	NM_020974	CEGP1	CEGP1 protein	Not annotated, but literature suggests secreted protein	1.327039456
208658_at	NM_004911	ERP70	protein disulfide isomerase related protein (calcium-binding protein, intestinal-related)	Endoplasmic reticulum	1.221133216
211048_s_at	NM_004911	ERP70	protein disulfide isomerase related protein (calcium-binding protein, intestinal-related)	Endoplasmic reticulum	1.263116031
210074_at	NM_001333	CTSL2	cathepsin L2	Lysosome	1.310239416
212290_at	AL050021		Homo sapiens mRNA; cDNA DKFZp564D016 (from clone DKFZp564D016)	Membrane protein	1.212390931
212295_s_at	AL050021		Homo sapiens mRNA; cDNA DKFZp564D016 (from clone DKFZp564D016)	Membrane protein	1.223423287
213094_at	AL080079	DKFZP564D0462	hypothetical protein DKFZp564D0462	Membrane protein	1.345110519
219410_at	NM_018004	FLJ10134	hypothetical protein FLJ10134	Membrane protein	1.209841065
221675_s_at	NM_020244	LOC56994	cholinephosphotransferase 1	Membrane protein	1.355728911
210512_s_at	NM_003376	VEGF	vascular endothelial growth factor	Membrane protein	1.086885938
211527_x_at	NM_003376	VEGF	vascular endothelial growth factor	Membrane protein	1.103813269
203988_s_at	NM_004480	FUT8	fucosyltransferase 8 (alpha (1,6) fucosyltransferase)	Membrane protein (by similarity).	1.206046136
203362_s_at	NM_002358	MAD2L1	MAD2 (mitotic arrest deficient, yeast, homolog)-like 1	Nucleus	1.111677987

Figure 1: **RNA content of fractions taken from the sucrose gradient.**
Vertical axis shows the absorbance at 260, and horizontal axis give the fraction number from the bottom of the gradient.

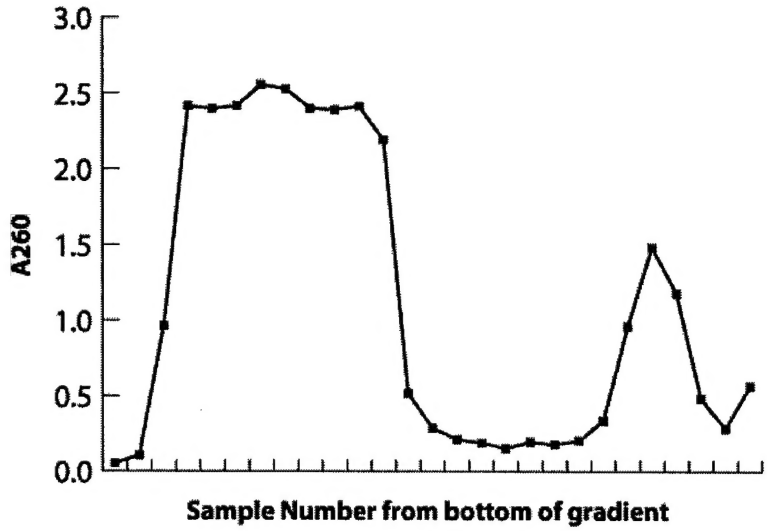

Figure 2: **Distribution of MEM/CYT ratios for all genes in the reference set expressing above the E_T.**

The vertical line indicates a MEM/CYT ratio of 1.08.

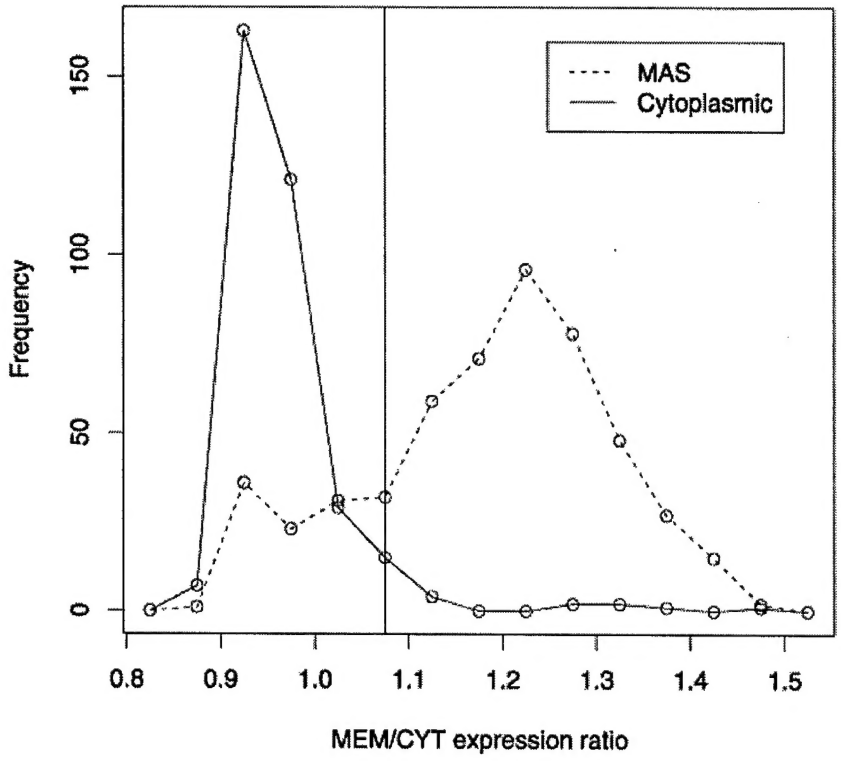

Figure 3: **Distribution of MEM/CYT ratios for genes which are not in the reference set expressing**

above the E_T. The vertical line indicates a MEM/CYT ratio of 1.08.